Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D468

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: April 7, 1997

Period of Report: January 1, 1997 to March 31, 1997

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
4/7/97

3. REPORT TYPE AND DATES COVERED
Scientific/Tech 1/1/97 - 3/31/97

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D468

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR528181397

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS
Browsing
Image Retrieval
Text Retrieval
Query Processing
Scanned Document Retrieval
Probabilistic Retrieval Model
Indexing
Bayesian Network
Large Distributed Databases

15. NUMBER OF PAGES
9

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we will be studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, and extending the underlying retrieval model to be able to make effective use of phrasal representations in both query-based retrieval and relevance feedback.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. Extensive use will be made of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query. This collection will be used for the experiments involving new probabilistic retrieval models and relevance feedback. Summarization techniques will be compared to sentence-based approaches and user-based evaluations of these summaries will be done. As more work is done on summarization in the TIPSTER program, we will make use of any new evaluation measures developed there.

Technical Results

We have continued investigating the benefits that can be derived from using phrases for query and document representation. For tasks that are specific to particular events in the news media (say), we have preliminary evidence that proper nouns and phrases (names, locations, companies) are of particular significance. We have started to investigate core concept identification in more detail, analyzing our results from TREC to determine where core concept analysis worked and where it failed. Simultaneously, we are investigating other query processing methods that rely less on natural language understand approaches and more on statistical methods (we continue to use both). We have begun looking into how representations of topics shift over time and how we can use that shifting to detect new sub-topics; this is being researched in the context of TREC-6 also. Our recent work on global measures of significance is being extended to non-Boolean operators.

Important Findings and Conclusions

Most of the work is not yet far enough along for conclusive results. We continue to analyze the query processing work in the context of TREC.

Significant Hardware Development

None

Special Comments

We worked with the PTO, San Diego Supercomputer consortium, and DARPA to acquire some of the patent collections. We were finally able to download some Greenbook data in very early April, after the period of this report. We are working on converting the data into a format which can be used by InQuery. We are also working on establishing a fast internet connection to aid this process. We understand it should be active by mid-May. We have also obtained some test data for classification work.

Implication for Further Research

We intend to apply our research to the patent data once we have appropriately converted it. We will then be able to refine our results based on feedback from PTO.

Task 2: Browsing and Discovery Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing collections of documents, and discovering connections between important ideas and documents in distributed collections. These techniques will be designed to support interactive browsing in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. In order to support discovery, connections must be made between documents and groups of phrases that use a variety of evidence in addition to direct co-occurrence.

General Methodology

The techniques will be evaluated with user-based and collection-based experiments. The relevance judgments from the TIPSTER collection will be used to evaluate clusters of documents. Phrase clusters will be evaluated by their impact on retrieval effectiveness and through user experiments that will measure performance on specific tasks. Part of the effort in this task (and the previous one) will involve developing a PTO test collection, which means that sample queries will need to be gathered from patent examiners and they will need to evaluate demonstrations of tools as they are developed.

Technical Results

Our 3-D graphics interface has been extended to include additional methods for document and term clustering and visualization. The clustering can now dynamically adjust as documents are interactively judged relevant (or non-relevant) to highlight unjudged documents' relationships. The document representation for clustering can also be selected dynamically. Some aspects of this work will be evaluated using the TREC interactive task.

Important Findings and Conclusions

A preliminary user study showed that a single-link clustering visual could be used by a user to predict significantly better clusters. However, better clustering methods were able to outperform the human (when the human started with weak clusters).

Significant Hardware Development

None

Special Comments

The PTO data is heavily fielded and we are working to understand which fields are useful and what their purposes are.

Implication for Further Research

Now that the PTO data is available to us (as of early April), we hope to apply some of these visualizations to that data. We expect that the structured data in the PTO texts (e.g., classification information) will play a key part in new visuals.

Task 3: Scanned Document Indexing and Retrieval

Task Objectives

The goals of this task are to develop techniques for detecting text, trademarks, logos, and images in scanned documents, clean up backgrounds of these detected objects, and support retrieval of images (such as designs in design patents), trademarks, and text from OCR.

Technical Problems

Current zoning techniques available with commercial OCR devices do not accurately locate text or trademarks within other images. We are developing techniques based on gaussian derivative filters to both detect and clean up (remove noisy backgrounds) these classes of objects in scanned documents. We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images and scanned documents. Specifically, we are working to obtain large collections of trademarks and design patents, as well as typical queries.

Technical Results

We continue to improve the indexing and retrieval techniques for images. We have focused in particular on improving our techniques for extracting text from images. We have applied the color-based retrieval on a different and larger database with successful results. We are working on a Java-based prototype to demonstrate our image retrieval work, and have ported the existing demo to the SGI platform.

Important Findings and Conclusions

The methods for extracting text from images are more robust and work more accurately. Color-based retrieval works on a larger set of sample images than we had originally tried it on.

Significant Hardware Development

None

Special Comments

We have acquired a subset of the trademark images and are converting them to a format we can use. We have also obtained about 900 patent image files which we are converting for use. A small number of plant patents have been scanned for study.

Implication for Further Research

We can now work with patent and trademark images.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

We have completed a preliminary study of InQuery's ability to scale to large databases. We have corrected some processing assumptions which interfered with efficient building of indexes, and are investigating some questions related to that process. We have also investigated techniques for rapidly retrieving documents from a very large collection without impacting effectiveness too much.

Important Findings and Conclusions

With some modest improvements, InQuery is capable of handling medium-sized collections (20 gigabytes). We are preparing to test larger collections.

Significant Hardware Development

The disk array purchased at the last report has arrived and been installed.

Special Comments

We are still waiting for the fast network access to PTO (anticipated mid-May) so that we can acquire more of the patent data for these large-collection tests.

Implications for Further Research

We will continue scale-up investigations, focusing also on collection selection and result merging.

PTO Quarterly Report
Distribution List

ARPA Agent: Harry Koch
ESC/AXS
Bldg 1704, Rm 114
5 Eglin Street
Hanscom AFB, MA 01831-2116

ARPA/ITO
ATTN: Gary Koob
3701 N Fairfax Drive
Arlington, VA 22203-1714

ARPA/Technical Library
3701 N Fairfax Drive
Arlington, VA 22203-6145

Defense Technical Information Center (DTIC)
Cameron Station
Alexandria, VA 22034-6145

ESC/ENK
ATTN: Ms Carole Stephan
Bldg 1704, Rm 119
5 Eglin Street
Hanscom AFB, MA 01731-2116
(Letter of Transmittal Only)

Lawrence J. Cogut
Office of System Architecture and Engineering
U.S. Department of Commerce
Patent and Trademark Office
Crystal Park 2, Suite 1004
2121 Crystal Drive
Arlington, VA 22202
ph. (703) 305-8685
fax (703) 305-9216
lcogut@uspto.gov